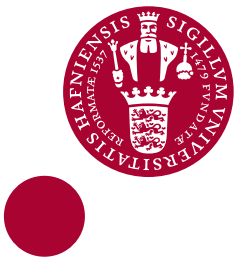

Domain specific MT in use

Offersgaard, Lene; Povlsen, Claus; Almsten, Lisbeth Kjeldgaard; Maegaard, Bente

Published in:
EAMT 2008: 12th annual conference of the European Association for Machine Translation

Publication date:
2008

Document version
Publisher's PDF, also known as Version of record

Domain specific MT in use

Lene Offersgaard[1], Claus Povlsen[1], Lisbeth Almsten[2], and Bente Maegaard[1]

[1]Center for Sprogteknologi, Københavns Universitet, Njalsgade 140, 2300 København S, Denmark
[2]Inter-set Translation,Vestergade 13A, 6500 Vojens, Denmark

{leneo, cpovlsen, bmaegaard}@hum.ku.dk, LKJ@intersetgroup.dk

Abstract. This paper focuses on domain specific use of MT with a special focus on SMT in the workflow of a Language Service Provider (LSP). We report on the feedback of post-editors using fluency/adequacy evaluation and the evaluation metric 'Usability', understood in this context as where users on a three point scale evaluate the sentence from the point of view of the post-editor. The post-editor profile defined by the LSP is based on the experiences of introducing MT in the LSP workflow. The relation between the Translation Edit Rate (TER) scores and 'Usability' scores is tested. We find TER a candidate for an automatic metric simulating the post-editors' usability judgements. LSP tests show 67% saved time in post-editing for the tested domain. Finally, the use of weighted sub-domain phrase tables in a SMT system is shown to improve translation quality.

Introduction

As part of a general strategy to strengthen cooperation between the research community and small and medium-sized enterprises, the Danish Council for Strategic Research has decided to co-finance two projects involving machine translation. The aim of both projects has been to explore the possibilities of using statistical machine translation (SMT) approaches in small and medium sized companies (SMEs). The project goals were to find out not only whether it would possible for translation companies to integrate SMT systems in their daily translation flow, but also to assess whether it would be financially beneficial. The primary tasks of the involved translation companies have been to provide bilingual corpora consisting of sentence aligned documents and then subsequently to test and evaluate the translation results as a first step to uncover the commercial potential of using SMT in their translation process.

This paper builds on the results of the second project, mainly carried out in 2007, involving at the research side Copenhagen Business School and the University of Copenhagen and on the business side Inter-Set, a medium-sized language service provider (LSP) with subsidiaries in other countries.

The Open Source Moses MT system [1] was used both for training the SMT system and for translation. MOSES is currently mainly supported under the EuroMatrix project, funded by the European Commission. The language model was

trained using the language modelling toolkit IRSTLM [2]. The language models were trained with order 5. The maximum length of phrases in the phrase tables was set to 5.

Domain Issues in SMT

The assumption that there would be a productivity gain using SMT was the prime motivation factor for the LSP to investigate the potential of SMT systems. One of the issues to be considered in this context was the handling of subject domains.

In an ideal world, all users involved with translation of technical documents would apply the same large-scale general subject classification system such as Lenoch [3]. From an SMT point of view the advantages of a consistent use of a classification system would be obvious. Not only would it ease the identification of consistent and representative bilingual training data, it would also, via the fine-grained subject classification, increase the probability that the lexical coverage of a given SMT-system would be tuned for the texts to be translated.

But unfortunately experiences show that use of a large universal classification system involves too much administrative work [4]. In addition, subject classification systems do not take into account possible divergences in the data within the same subject domain, e.g. different companies may have chosen to use different specific company terminologies.

Besides, texts from the same subject domain will make use of very different writing styles in terms of sentence types and varieties in language usage according to the genre of the text. Marketing texts, for instance, may praise the features of the product while manuals focus on strict instructions on how to use the product.

Consequently, in principle it would be preferable to train an SMT system on texts with almost identical writing styles and within the same subject domain. On the other hand, for practical and financial reasons, it would desirable that the SMT system had a broad coverage being usable for different text types without a negative impact on the translation quality. So, the solution is a compromise.

In general the LSPs are very aware that different products and clients use different writing styles in terms of sentence types and variation in language usage and terminology. With a focus on delivering high quality translation, it is obvious that the clients' expectations regarding correct handling of terminology and writing style had to be met, also for the SMT system. Therefore, the researchers and the LSP in collaboration tested the suitability of 5 different sub-domains of manuals to see how well these different sub-domains could be translated with the SMT system.

Selection of Sub-domains

The LSP chose 5 candidate sub-domains within the domain of technical manuals and collected training material for these topics. The training data were aligned sentences extracted from the LSP's translation memories (TMs) and consisted of approximately 135,000 parallel sentences. From each sub-domain a development test corpus of 250 test sentences was extracted.

Table 1. Training and test material for English->Danish SMT system.

Sub-domains	Training words (Danish)	Training sentences	Training sentence av. length	Development test set Words	Development test set Sentences
A:Camcorders	262,138	24,897	10.5	3,263	250
B:Software	1,609,943	73,517	21.9	7,282	250
C:DVD	136,683	12,991	10.5	2,416	250
D:Printers	144,379	14,657	9.9	1,989	250
E:Mobile phones	127,701	8,740	14.6	3,216	250
Total	2,280,844	134,802	16.9	18,166	1,250

To do a quick evaluation of the system it was decided to score the translation output with the two automatic metrics BLEU [5] and TER [6]. As can be seen the sub-domain with the best score is B, the second best score is found for sub-domain E, where a high figure for BLEU and a low figure for TER state that the translation output is closer to the reference, cf. table 2 below.

Table 2. Translation quality in terms of BLEU and TER scores for 5 sub-domains.

Development test data:	BLEU	TER
A:Camcorders	0.5517	33.17
B:Software	0.7564	16.81
C:DVD	0.4766	37.69
D:Printers	0.6539	23.71
E:Mobile phones	0.6713	24.82
Total	0.6818	24.72

Based on this evaluation, the sub-domains B, E and D seemed to the best candidates to focus on. For sub-domain B, however, the client had already started to use MT and an additional SMT system for this sub-domain therefore turned out to be without interest. The LSP chose to focus on sub-domain E, since sub-domain D has shorter sentences and better match in TM.

Evaluation Test Results

The focus point of MT evaluation differs dependent on your perspective. From the developers' point of view, evaluation as part of testing the MT system has to be quick and cheap. While from the users' point of view, the evaluation has to focus on easier use, better translation quality, quicker post-editing etc. We have carried out both types of evaluation and compared results.

Automatic evaluation measures

For the system developer and for system tuning the goal is to have automatic metrics that give reproducible results and save users from doing expensive post-editing tasks in each iteration of system improvements. Two automatic evaluation metrics were used: BLEU metric [5] and Translation Edit Rate, TER [6]

There has been much focus on evaluation of SMT and MT-systems in the last decades [7], [8], [9]. For a brief overview of other currently used evaluation metrics used for SMT and MT and recent experiences within the field, see [10] and [11]. The two selected metrics were chosen because they are easy for the developer to apply, given a translation reference. It has been argued that an increase/decrease in the value of the BLEU score does not guarantee a better/worse translation quality [13]. But nevertheless, the metric is still widely used to measure development improvements in systems. TER is calculated as the ratio of edits (insertions, deletions and substitutions of single words as well as shifts of word sequences) compared to the average number of words in the references. TER is stated to correlate reasonably well with human judgements [6]. TER values will be in the range from 0 (translated sentence is exactly like the reference) to in principle more that 100, e.g. if the reference sentence consists of only a few words whereas the translation output contains too many words and therefore needs more edits that the length of the reference sentence.

User evaluation measures

From a user's point of view, automatic evaluation figures are somewhat abstract and difficult to comprehend and do not necessarily provide feedback to the questions raised above. Alternative evaluation metrics focussing much more on the human translation aspect have conceived in order to meet this problem. The following metrics represent this alternative evaluation approach:

- Fluency and adequacy scoring
- Usability scoring
- Post-editing time

Fluency and adequacy have been defined using a five point scale [8]. Recent studies show that scores for fluency and adequacy apparently do not correlate very well between users[1], and therefore these score results would be more difficult to use for system testing and tuning. The LSP post-editors involved in evaluating SMT output stated that a five point scale would be much too difficult to use. We therefore reduced the scale to a four point scale which gave the users an easier job and thereby probably more reliable results.

The measure that the users suggested themselves is here called Usability. In an LSP environment, the conventional translation platform is a TM platform. 'Usability' is a measure that allows post-editors to score a machine-translated translation unit in terms of usability compared to a fuzzy match in a TM tool. A machine-translated translation unit may not be adequate or fluent, but it may be usable. When it is usable, the time needed to edit the machine-translated translation unit will be shorter than the time needed to translate the segment from scratch. It is in this context defined as a three point scale. The scale is focused on the post-editing process, and the user can use the following scores:

3: Good translation – few key strokes needed to edit translation. Corrections of casing or layout can be needed. Use of terminology is correct.

2: Translation can be post-edited using less time than a translation of the sentence from scratch – number of key strokes needed to edit translation is less than the key strokes needed to translate from scratch.

[1] Koehn, Philipp. Invited talk MT SUMMIT XI, Copenhagen 2007

1: Translation quality is too poor. It will take more time to post-edit the sentence, than to translate the sentence from the source sentence – translation is discarded.

The post-editors at the LSP found this scoring very useful as it is closely connected to their translation workflow, no matter whether they use TM as their translation tool or post-edit MT-output.

The judgement tool for usability, adequacy and fluency shows the source sentence and the translation output. The time used to do the scoring of all three measures is also measured. The goal of doing user evaluation is first of all to give the users tools to evaluate the SMT output in order to give feedback to system development and secondly to compare the judgements with the figures from automatic evaluation. As human evaluation is costly, we present the results we got in all the evaluation tasks although the amount of data is small. The test involved three in-house post-editors. All of them were experienced proof-readers. Two of them were domain specialists. One post-editor was a general quality assurance specialist.

In table 3 the average scores for usability, adequacy and fluency together with the BLEU and TER scores are given. The BLEU and TER scores are not as good for the test sets as for the development test set.

Table 3. The human average scores and the automatically computed scores. Usability scale 1-3, adequacy scale 1-4 and fluency scale 1-4.

Test set	Sentences	Average score Usability	Average score Adequacy	Average score Fluency	BLEU	TER
A	732	1.94	2.30	2.18	0.57	32.1
B	117	2.02	2.35	1.89	0.51	34.6
Development test set	1250				0.68	24.3

We also split up the fluency and adequacy scores in relation to the usability scores (Table 4). A clear correlation between the fluency/adequacy scores and the usability scores can be seen. The users' feedback on the fluency and adequacy scoring was that they would prefer only to use the usability score, as it is a simple measure. In addition, it would be easier for the post-editor to do this kind of judgement because this is the way that they normally conduct translation quality assessments. They also stressed that it is important that the scoring can be done fast for each sentence, preferably less than 30 sec. on average. As can be seen the scores are done in 14 sec. on average for sentences with usability=1, and in 20 sec. for usability=3.

To compare the user evaluation scorings with the scorings done by automatic metrics, we focus on the TER metric. The relation between the TER scoring and the usability score are given in Table 5. It can be seen that when considering individual sentences there are significant noise on the TER scoring because the standard deviation is approximately as large as the distance between the usability classes. However, when considering a text with n sentences the noise on the overall TER scoring decreases by a factor $n^{-½}$. Hence, if we have 100 sentences the standard deviation on the overall TER scoring corresponds to approximately 0.1 usability units. Therefore we consider TER a promising candidate for an automatic metric for simulating the post-editors' judgements on text level. Feedback from the post-editors

also mention that a very good automatic metric has to take into account that words from some word classes are more important to translate correctly than others.

Table 4. Relation between usability and fluency/adequacy scorings. Average time for scoring is shown depending on usability scoring, but covers the time used giving all three scorings.

	Fluency				Adequacy				Time, average
Usability	**4**	**3**	**2**	**1**	**4**	**3**	**2**	**1**	**sec.**
3	99%	1%			100%				14.1
2		10%	90%			28%	72%		17.4
1			5%	95%			6%	94%	19.8

Table 5. Relation between usability and TER. For all sentences in each usability class the average and the standard deviation of the TER scoring are given.

	TER	
Usability	**Average value**	**Standard deviation**
3	12.0	18.1
2	34.4	24.9
1	52.9	23.3

Post-editing Experiences in a 'Real-life' Translation Project

Shortly after the LSP had joined the project, the LSP received a commercial order for a translation project involving machine translation. It was a Microsoft (sub-domain B) project in which the LSP's job was to post-edit already machine-translated text strings instead of translating text in a TM environment.

Post-editing was a more complex job than they had expected. In many ways, post-editing is comparable to proof-reading human translations, however in some ways it differs. The LSP set out by hiring experienced Microsoft translators to do the post-editing. Unexpectedly, when delivering the first post-editing assignment, it did not pass Microsoft's validation process so they learned that good translators are not necessarily good post-editors.

Another aspect of the post-editing project was that rates were lower than the rates of ordinary translation projects. Low rates are based on the assumption that there is a productivity gain when using machine translation compared to human translation in a TM environment.

Based on the lessons learned from this Microsoft project it can be concluded that good post-editing skills differ from good translation skills so other vendor profiles have to be chosen and secondly that low rates on post-editing require that less time can be spent on translation.

Post-editor Profile

What is the ideal post-editor profile then? First of all, the ideal post-editor has to fulfil the requirements stated by clients in the localization industry. These requirements include:

- The use of consistent terminology
- Continuity with existing translations => existing translated text strings cannot be altered
- Compliance with client's style guides
- Ability to observe deadlines
- Ability to adapt level of quality to price pressure restraints

In our experience, this is the ideal post-editor profile:

- A good post-editor knows the domain
- A good post-editor has very good skills in his/her native language
- A good post-editor can implement style guides and consistent terminology.
- A good post-editor is an experienced proof-reader. He/she can make swift decisions on "good – no good" in order to be able to discard machine-translated text strings that are not worthwhile post-editing, but need to be translated from scratch.

In many ways, the profile of a good post-editor fits a good translator. However, there is one significant and very important difference: the ability to decide – within a few seconds - whether a translation should be discarded. Many translators tend to spend too much time on this decision process.

Comparing SMT with TMs

What really matters for the users is the amount of time spent on post-editing the output, compared with the amount of time spent on translation with the tools they are used to in the translation process. As an LSP, the TM environment is the standard translation platform. Consequently, one of Inter-Set's goals in the project was to compare SMT with TM. Also, based on the low post-editing rates compared to standard translation rates, it was – from a financial point of view – interesting to see whether using SMT did actually produce a productivity gain compared to the traditional TM environment.

In the domain of mobile phones (sub-domain E), a TM translation was compared to an SMT translation. First the text to be translated in the TM tool was analysed. The result of the analysis can be seen in table 6. Then it was estimated how much time would be needed to translate the text in a TM environment.

It turned out that translating the remaining 1,749 words in a TM environment took 6 hours, while the translation task using SMT took 2 hours giving a productivity gain of 67%. Even though the amount of test data is sparse, this investigation indicates that a productivity gain can be used shifting from a TM environment to an SMT system.

We also investigated the connection between the TM match percentages and the Usability scores of the two test documents. Table 7 shows that in the two tests 16% (A) and 8% (B) have a match of 95%-100% in the TM, but a larger percentage of the sentences get the score usability=3 (Good translation): 17% and 20%. It can also be seen that 67% and 81% of the sentences have “No match” in the TMs. Compared with the Usability scores only 23% and 19% of the sentences get the score Usablity=1. This also indicates that using SMT for this text type could be beneficial.

Table 6. Analysis of text B to be translated in TM tool.

Match Types	Segments	Words	Percent	Placeables
Context TM	0	0	0	0
Repetitions	2	2	0	0
100%	3	36	2	0
95% - 99%	0	0	0	0
85% - 94%	3	24	1	0
75% - 84%	7	84	5	0
50% - 74%	4	39	2	0
No Match	161	1,600	90	0
Total	180	1,785	100	0
Chars/Word	4.59			
Chars Total	8,204			

Table 7. Comparison of sentences (in %) split up by TM match-% and Usability scores.

TM matches	Test A %sentences	Test B %sentences	Usability score	Test A %sentences	Test B %sentences
95%-100% ”3”	16	8	Usability=3	17	20
50%-94% ”2”	17	11	Usability=2	59	61
”No match” ”1”	67	81	Usability=1	23	19
”Average”	1.49	1.25	Average	1.94	2.02

Use of Sub-domain Phrase Tables

The use of sub-domain phrase tables addresses the point that the LSP would like to gain as much as possible from the client/text type specific training data, but on the other hand also would like to have a broad coverage of words and phrases. Training an SMT system on a small amount of training material for a given sub-domain leads to a narrow lexical coverage which again results in low translation quality. In order to cope with this problem and to get a better system performance a number of phrase tables were combined. For the training material available (Table 1) four out of five sub-domain corpora are very small (less than 300,000 words for each sub-domain). As expected we got poor results when translating the development test set based only on these small sub-corpora.

Consequently we decided to combine the sub-domain phrase tables with the phrase table based on all the five sub-domain corpora. This combination can be done using the MOSES decoder as MOSES has an option allowing more phrase tables to be used in the same translation process. In this test the sub-domain specific translation table is given a weight of 0.5 while the general phrase table is given a weight of 0.2. The results of translating the five sub-domain test corpora compared with the results using only one phrase table is shown in Table 8.

Table 8. BLEU and TER scores for sub domain test sets translated using one common phrase table and translated with two weighted phrase tables.

Sub domains	BLEU one phrase table	BLEU two weighted phrase tables	Improve-ment	TER one phrase table	TER Two weighted phrase tables	Im-prove-ment
A	0.5517	0.5894	0.0377	33.17	30.61	2.56
B	0.7564	0.7785	0.0221	16.81	15.01	1.80
C	0.4766	0.5133	0.0367	37.69	34.92	2.77
D	0.6539	0.6530	-0.0009	23.71	23.61	0.10
E	0.6713	0.6986	0.0273	24.82	22.94	1.88
Total	0.6818	0.6870	0.0377	24.72	22.83	1.89

It can be seen that for the sub-domains A, B, C and E both the BLEU and the TER scores are significantly better yielding an increase of 2.2 to 3.7 BLEU points and of 1.8 to 2.8 TER points. We find it likely that further improvement can be obtained by such a type of optimization and regard it as a promising strategy for at the same time to give sub-domain training material priority and obtain a broader lexical coverage. In a study done by [12] test of domain corpus adaptation for broader domain coverage is reported with an increase of 1.7 BLEU points. Combinations of domain corpora with corpora covering general language are under further investigation in collaboration with the LSP, including weighting of the phrase tables that favours the use of the domain terminology rather than the general vocabulary.

Conclusion

We have described the post-editor profile based on experiences conducted in connection with the introduction of SMT in the translation workflow of an LSP. The LSP post-editors preferred the evaluation metric 'Usability' rather than fluency/adequacy, as it is more closely related to their every-day workflow when evaluating translations. The automatic metric TER was tested as a candidate for an automatic metric with correlation to human judgement, and the TER appeared to be a promising candidate.

With a productivity gain of 67 % saved time in post-editing in the test of SMT, the obvious next step for the LSP would be to integrate SMT in their workflow for some domains. Gaining benefits of machine translation however depends on the data that are available and used for the training of the SMT engine.

TMs may be client-specific including all text types, from user manuals to meeting notes. In these TMs, terminology may be consistent, but sentence structures may differ widely. For an LSP with many clients and many rather small client-specific TMs (each containing a few hundreds of thousands of words), the target of millions of words for the training of an SMT engine is simply out of reach. Here the combination of weighted phrase tables looks as a promising strategy to overcome this threshold.

Acknowledgements

We would like to thank Philip Koehn, Edinburgh University for providing a judgement tool that we could modify and use in the tests.

References

1. Koehn, Philip, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello Bertoldi, Nicola Cowan, Brooke, Shen, Wade, Moran, Christine, Zens, Richard, Dyer, Chris Bojar, Ondrej Constantin, Alexandra and Herbst, Evan: Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL), Prague, Czech Republic (2007)
2. Federico, Marcello and Cettolo, Mauro: Efficient Handling of N-gram Language Models. for Statistical Machine Translation. In Proceedings of the Second Workshop on Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL), Prague, Czech Republic (2007)
3. Wellenstein, H. (upd.): Lenoch Universal Classification System, LUC – vol. 2, Main part. 2. edition. European Commission, Translation Service – Terminology and Language Support Services (1996)
4. Rirdance, S., Vasiļjevs. A.: "Towards Consolidation of European Terminology Resources. Experience and Recommendations from EuroTermBank Project. Riga, Latvia (2006)
5. Papineni, Kishore, Roukos, Salim, Ward, Todd and Zhu Wei-Jing: BLEU: A method for automatic evaluation of machine translation. In Procedings of ACL (2002)
6. Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," Proceedings of Association for Machine Translation in the Americas (2006)
7. Van Slype, G.: Critical Study of Methods for Evaluating the Quality of Machine Translation. Final Report, Bureau Marcel van Dijk / European Commission, Brussels (1979)
8. White, John and O'Connell, Theresa A.: Evaluation in the ARPA Machine Translation Program: 1993 Methodology. ACL (1994)
9. LDC 2005, Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Revision 1.5. (2005)
10. Callison-Burch, Chris and Fordyce, Cameron and Koehn, Philipp and Monz, Christof and Schroeder, Josh: (Meta-) Evaluation of Machine Translation. In Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic, Association for Computational Linguistics, pp. 136-158. (2007)
11. Callison-Burch, Chris and Fordyce, Cameron and Koehn, Philipp and Monz, Christof and Schroeder, Josh: Further Meta-Evaluation of Machine Translation. In Proceedings of the Third Workshop on Statistical Machine Translation, Columbus, Ohio, ACL, pp.70-106, http://www.aclweb.org/anthology/W/W08/W08-0309. (2008)
12. Koehn, Philip and Schroeder, Josh: Experiments in Domain Adaptation for Statistical Machine Translation. In Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic, Association for Computational Linguistics, pp. 224-227. (2007)
13. Callison-Burch, Chris, Miles Osborne and Philipp Koehn: Re-evaluating the Role of Bleu in Machine Translation Research. In Proceedings of EACL (2006)